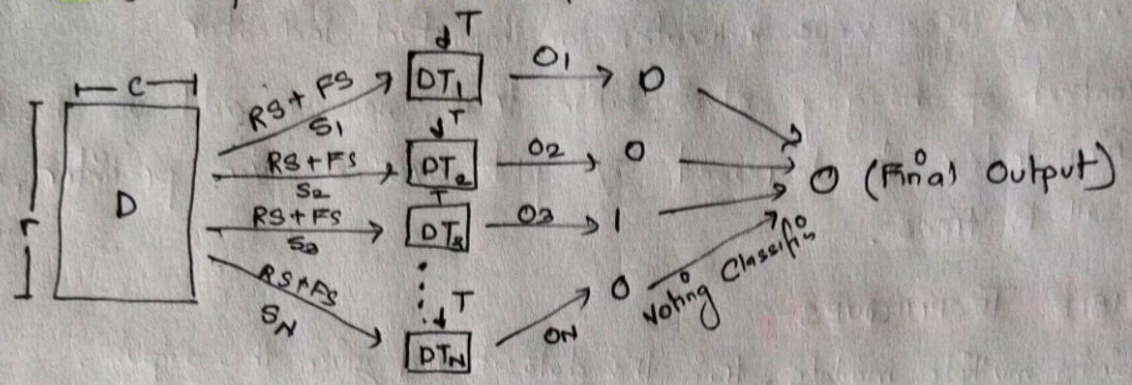
Random forest

- It is a bagging technique.
- RS → Row sampling with replacement, Feature sampling with replacement → FS
  Dataset → D, Decision tree → DT, columns/feature → C, rows → r
  Model → m, Output → O., Sample → S,

Eg



- Here base learner is decision tree.
- Decision tree is low Bias, high variance. Low bias means low training error and high bias means high test error. So if we build decision tree to its fullest, it overfit the data.
- Random forest is a collection of decision tree which will make a model, final model into low bias and low variance.
- So how it acheive low variance?
  In collection of decision tree, each decision tree become expert in certain rows and features because of row sampling with replacement and feature sampling with replacement. And we choose the output based on voting classifier which will choose the output based on highest vote received. In short, it become a generalised output means it have low variance.
- So random forest have low bias, low variance and generalised output.
- In regression, all the model will give continuous value. So random forest take the mean or median of all the continuous value and give it to one value. In sklearn, it give the average or mean of all the continuous value.

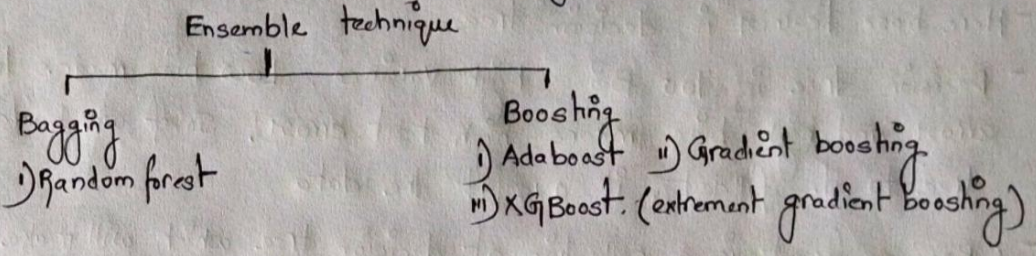
(13)

Variable Importance Measure –

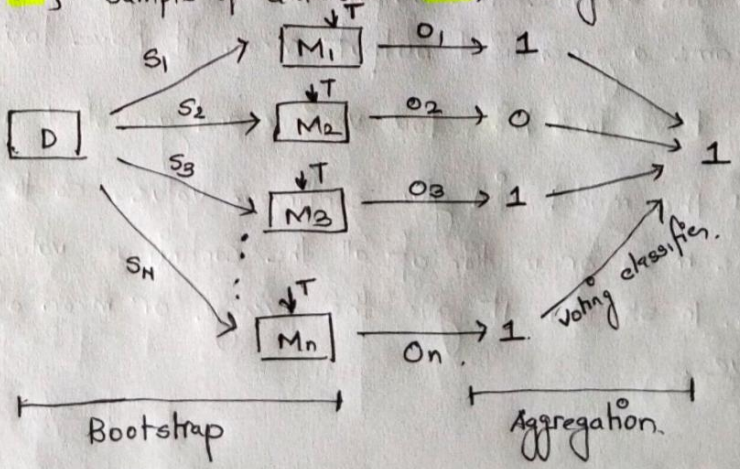
- One can obtain an overall summary of the importance of each predictor using the RSS (for bagging regression tree) or Gini Index (for bagging classification tree)
- In case of bagging regression tree, we can record the total amount that the RSS is decreased due to splits over a given predictor, averaged all over B trees. A large value indicates an important indicator.
- In context of bagging classification tree, we can add up the total amount of Gini Index is decreased by split over a given predictor, averaged over all B trees.

ENSEMBLE TECHNIQUES –

- Combining multiple models, train the dataset and given the output.



- Bagging, also known as bootstrap aggregation.
- Row sampling with replacement technique is used in bagging. Suppose we have to 5 different set of samples. So first set is choosen by taking some random sample. Next set is choosen taking random sample including the sample taken in first set. That is row sampling with replacement technique.
- Dataset → D, Sample of dataset → S, Voting classifier → V, Model → M, Baselearners. Output → O. Test data → T



Bootstrap method is resampling technique used to estimate statistics on a population by sampling a dataset with a replacement.

- In bagging we have different Models $M_1, M_2, M_3 \ldots, M_n$. Dataset D, is divided into samples with help of row sampling with replacement. Then each model is train on samples and then we give test data to each model, and model genrated output $O_1, O_2, \ldots, O_N$. Based on voting classifier, the output which got most vote, is selected as final output.